Week 16b

Last Time

- Areal Data Visualization
- Assessing Spatial Structure in Areal Data
- Overview of Areal Data Models

This Time

- Model fitting with Areal Data
- Simulating the spatially correlated areal data
- Modeling continuous spatially correlated areal data

Recall: Disease Mapping

Areal data with counts is often associated with disease mapping, where there are two quantities for each areal unit: Y_i = observed number of cases of disease in county i and E_i = expected number of cases of disease in county i. Generically, this can be modeled as $Y_i|\psi_i \sim Poisson(E_i\psi_i)$.

We will use `S.glm`, `S.CARbym`, and `S.CARleroux` from the `CARBayes` package to fit and compare models using deviance information criteria.

```
formula <- observed ~ offset(log(expected))
no_spatial <- S.glm(formula=formula, data=respiratory_admissions,
                          family="poisson", burnin=10000, n.sample=30000,thin=2, verbose =

print(no_spatial)
```

```
#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - None
Regression equation - observed ~ offset(log(expected))


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 10000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 2

############
#### Results
############
Posterior quantities and DIC

       Mean       2.5%      97.5% n.effective Geweke.diag
    -0.1645    -0.1839    -0.1448   4794.3000      0.0000

DIC =  2288.324       p.d =  1.036617       LMPL =  -1149.3
```

```
exp(-.1643)
```

```
[1] 0.8484874
```

```
mean(respiratory_admissions$SMR)
```

```
[1] 0.8605064
```

One way to incorporate spatial structure is with the Besag-York-Mollie (BYM) model, written as

$$
\begin{aligned}
Y_i|\psi_i &\sim Poisson(E_i\psi_i) \\
log(\psi_i) &= x_i^T\beta + \theta_i + \phi_i
\end{aligned}
$$

where we place a CAR prior on ϕ and standard random effects on θ.

$$\begin{aligned}\phi_k|\phi_{-k}, W, \tau &\sim N(\frac{\sum_{i=1}^k w_{ki}\phi_i}{\sum_{i=1}^k w_{ki}}, \frac{\tau^2}{\sum_{i=1}^k w_{ki}}) \\ \theta_k &\sim N(0, \sigma^2)\end{aligned}$$

```
bym <- S.CARbym(formula=formula, data=respiratory_admissions,
                             family="poisson", W=W_mat, burnin=10000,
                n.sample=30000, thin=2, verbose = F)
print(bym)
```

```
#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - BYM CAR
Regression equation - observed ~ offset(log(expected))


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 10000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 2

############
#### Results
############
Posterior quantities and DIC

               Mean    2.5%   97.5% n.effective Geweke.diag
(Intercept) -0.2204 -0.2433 -0.1968      5144.8        -1.5
tau2         0.3699  0.1828  0.5404        83.9         0.3
sigma2       0.0156  0.0026  0.0557        45.7        -0.4

DIC =  1072.668       p.d =  115.8446       LMPL =  -584.79
```

Alternatively we can specify the following model known as the Leroux model which uses the IAR framework with the ρ term where

$$\begin{aligned}
Y_i|\psi_i &\sim Poisson(E_i\psi_i) \\
log(\psi_i) &= x_i^T\beta + \phi_i \\
\phi_k|\phi_{-k}, W, \tau &\sim N(\frac{\rho\sum_{i=1}^k w_{ki}\phi_i}{\rho\sum_{i=1}^k w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{i=1}^k w_{ki} + 1 - \rho})
\end{aligned}$$

```r
leroux <- S.CARleroux(formula=formula, data=respiratory_admissions,
                             family="poisson", W=W_mat, burnin=10000,
                n.sample=30000, thin=2, verbose = F)
print(leroux)
```

```
#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - Leroux CAR
Regression equation - observed ~ offset(log(expected))


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 10000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 2

############
#### Results
############
Posterior quantities and DIC

               Mean    2.5%   97.5% n.effective Geweke.diag
(Intercept) -0.2208 -0.2451 -0.1976      3623.5         0.6
tau2         0.3366  0.2214  0.4791      2685.0        -0.7
rho          0.6239  0.3176  0.9180      1730.9        -0.3

DIC =  1073.271       p.d =  116.707       LMPL =  -581.93
```

Note that the above models result in a single smooth, spatial random surface (defined by the neighborhood structure). The differences in the BYM and the Leroux approaches are fairly minimal.

However, models can also be formulated to incorporate local spatial structure.

One option is the Lee and Mitchell approach, which models the w_{kj} terms rather than setting all to be zero or one. Specifically, an additional variable (Z) is constructed to model dissimilarity between neighboring units. In this case, our z values correspond to the percentage of people defined to be income deprived. Using this value we construct a distance (or dissimilarity) metric between areal units.

Fit this model using `S.CARdissimilarity` and compare to the previous models.

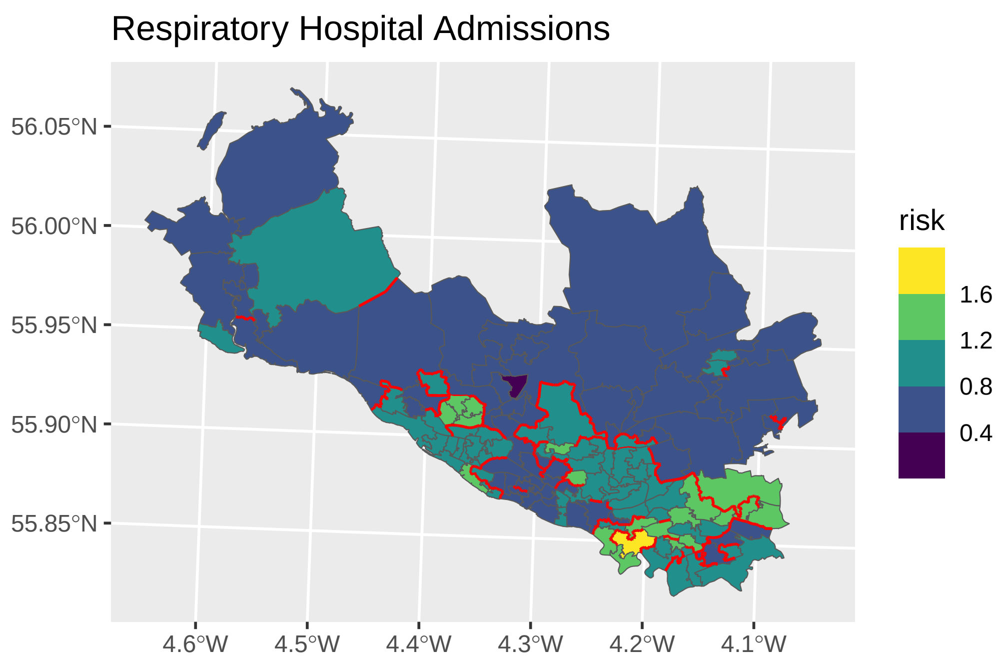
```
income <- respiratory_admissions$incomedep
Z.incomedep <- as.matrix(dist(income, diag=TRUE, upper=TRUE))

dis <- S.CARdissimilarity(formula=formula, data=respiratory_admissions,
                          family="poisson", W=W_mat, Z=list(Z.incomedep=Z.incomedep), verl
                          W.binary=TRUE, burnin=10000, n.sample=30000, thin=2)

print(dis)
```

```
#################
#### Model fitted
#################
Likelihood model - Poisson (log link function)
Random effects model - Binary dissimilarity CAR
Dissimilarity metrics -  Z.incomedep
Regression equation - observed ~ offset(log(expected))


#################
#### MCMC details
#################
Total number of post burnin and thinned MCMC samples generated - 10000
Number of MCMC chains used - 1
Length of the burnin period used for each chain - 10000
Amount of thinning used - 2

############
#### Results
```

```
############
Posterior quantities and DIC

               Mean    2.5%   97.5% n.effective Geweke.diag alpha.min
(Intercept) -0.2196 -0.2416 -0.1979      4250.5         0.0        NA
tau2         0.1374  0.0969  0.1895      2985.0        -0.4        NA
Z.incomedep  0.0499  0.0466  0.0513      2520.6         0.1    0.0139

DIC =  1057.446       p.d =  98.77812       LMPL =  -556.55

The number of stepchanges identified in the random effect surface
     no stepchange stepchange
[1,]           261         99
```

We can also extract the boundaries, where a stepchange (no neighbor structure) is identified.

```r
border.locations <- dis$localised.structure$W.posterior
respiratory_admissions$risk <- dis$fitted.values /
  respiratory_admissions$expected
boundary.final <- highlight.borders(border.locations=border.locations,
                                sfdata=respiratory_admissions)
st_crs(boundary.final) <- raster::crs(respiratory_admissions)

respiratory_admissions |>
  ggplot() +
  geom_sf(aes(fill = risk)) +
  geom_sf(data = boundary.final, color = 'red') +
  scale_fill_viridis_b() +
  ggtitle('Respiratory Hospital Admissions')
```



Models for continuous data

Now consider a continuous response on areal data. We will use a dataset called `pricedata` on the same areal locations as our previous analysis.

```r
library(CARBayesdata)
data(pricedata)
head(pricedata)
```

```
         IZ   price crime rooms sales driveshop     type
1 S02000260 112.250   390     3    68       1.2     flat
2 S02000261 156.875   116     5    26       2.0     semi
3 S02000262 178.111   196     5    34       1.7     semi
4 S02000263 249.725   146     5    80       1.5 detached
5 S02000264 174.500   288     4    60       0.8     semi
6 S02000265 163.521   342     4    24       2.5     semi
```

```r
pricedata <- pricedata |>
  mutate(log_price = log(price))
```

Here is a data dictionary for this dataset:

- **IZ:** The unique identifier for each IZ.
- **price:** Median property price.
- **log_price:** We've created the logarithm of price, which can be useful for modeling given the skewed structure of price.
- **crime:** The crime rate (number of crimes per 10,000 people).
- **rooms:** The median number of rooms in a property.
- **sales:** The percentage of properties that sold in a year.
- **driveshop:** The average time taken to drive to a shopping centre in minutes.
- **type:** The predominant property type with levels: detached, flat, semi, terrace.

Note that the data curators deleted one observation due to an aberrant value.

Explore mean structure in `log_price` as a function of other variables with data visualization.

Visualize log price and assess spatial structure

Implement a statistical model for log price that includes spatial correlation